CEL0: a continuous alternative to l0 penalty

Emmanuel Soubies, Laure Blanc-Féraud, Gilles Aubert

CEL0: a continuous alternative to ℓ_0 penalty.

Emmanuel Soubies and Laure Blanc-Féraud
UNS, CNRS I3S, UMR 7271, Sophia Antipolis, France.
{soubies,blancf}@i3s.unice.fr

Gilles Aubert
UNS, CNRS J.A.D, UMR 7351, Nice, France.
gaubert@unice.fr

***Abstract*—This paper presents a new way to address the NP-hard combinatorial ℓ_2-ℓ_0 problem by minimizing a *continuous* relaxed functional *preserving the minimizers* of the initial energy. We propose the *Continuous Exact* ℓ_0 *penalty* (CEL0), an approximation of the ℓ_0 norm leading to a *tight continuous relaxation of the* ℓ_2-ℓ_0 *criteria whose global minimizers contain those of* the ℓ_0 penalized least-squares functional. Links between local minimizers of these two functionals are also investigated. This short communication summarizes the main results of our recent work [1].**

I. Context

In this work we deal with the following ℓ_0 penalized least squares problem

$$\hat{\mathrm{x}} = \arg\min_{\mathrm{x}\in\mathbb{R}^N} \mathrm{G}_{\ell_0}(\mathrm{x}) := \frac{1}{2}\|A\mathrm{x}-d\|_2^2 + \lambda\|\mathrm{x}\|_0\,, \qquad (1)$$

where $A \in \mathbb{R}^{M\times N}$, $M \ll N$, $d \in \mathbb{R}^M$, $\|\cdot\|_0$ denotes the ℓ_0-counting "norm" and $\lambda > 0$ is an hyperparameter allowing a trade-off between data fidelity and sparsity. This NP-hard combinatorial problem and its constrained form (not equivalent) are of fundamental importance in many applications such as coding, compressed sensing, source separation or variable selection.

During the last decades, many researchers proposed methods and algorithms, with some theoretical guaranties in terms of convergence, to find a good approximate solution of (1). The most popular are the ℓ_1 *convex relaxation* [2], the *greedy algorithms* [3] and the *continuous nonsmooth nonconvex penalties* widely used to replace the ℓ_0-norm in (1). Particularly, some authors have recently proposed an exact reformulation of ℓ_0 regularized problems as DC programs [4]. The following work comes within this framework of exact reformulation were a new penalty, called *Continuous Exact* ℓ_0 (CEL0), is proposed. This penalty leads to a *tight continuous relaxation* preserving the minimizers of G_{ℓ_0}.

II. Main contributions

Let us first introduce some notations used in the sequel:

- $\mathbb{I}_N = \{1, \cdots, N\}$,
- $a_i \in \mathbb{R}^M$ denotes the ith column of $A \in \mathbb{R}^{M\times N}$,
- $\sigma(\mathrm{x}) := \{i \in \mathbb{I}_N;\, \mathrm{x}_i \neq 0\}$ defines the support of $\mathrm{x} \in \mathbb{R}^N$,
- $\sigma^-(\mathrm{x}) := \left\{i \in \sigma(\mathrm{x}) : |\mathrm{x}_i| < \sqrt{2\lambda}/\|a_i\|\right\}$ a part of the support.

A. Orthogonal case analysis

When the matrix A is orthogonal, one can compute analytically the biconjugate $\mathrm{G}_{\ell_0}^{\star\star}$ which is the *convex hull* of G_{ℓ_0}. Simple calculations [1, §3] lead to,

$$\mathrm{G}_{\ell_0}^{\star\star}(\mathrm{x}) = \frac{1}{2}\|A\mathrm{x}-d\|^2 + \Phi_{\text{CEL0}}(\mathrm{x})\,, \qquad (2)$$

where Φ_{CEL0} denotes the CEL0 penalty defined by

$$\Phi_{\text{CEL0}}(\mathrm{x}) := N\lambda - \sum_{i\in\mathbb{I}_N} \frac{\|a_i\|^2}{2}\left(|\mathrm{x}_i| - \frac{\sqrt{2\lambda}}{\|a_i\|}\right)^2 \mathbb{1}_{\{|\mathrm{x}_i|\leq\frac{\sqrt{2\lambda}}{\|a_i\|}\}}\,. \qquad (3)$$

Therefore, in the case of orthogonal matrices, the convex hull of G_{ℓ_0} is obtained by replacing the noncontinuous ℓ_0-norm by the *continuous* CEL0 penalty given in (3). In this case, since $\mathrm{G}_{\ell_0}^{\star\star}$ is convex, all minimizers are global and we can deduce from them the sparsest solution of (1) which is given by thresholding the entries of $A^T d$ [1, §4.3]. The resulting thresholding rule can be seen as a generalization of the well-known hard thresholding rule [5].

However this result is false when A is not orthogonal. Indeed, in this case, replacing the ℓ_0 norm in (1) by the CEL0 penalty (3) leads to a nonconvex functional denoted G_{CEL0}. Nevertheless, this functional has interesting properties which are analyzed in the following.

B. Links between minimizers of G_{CEL0} and G_{ℓ_0}

Let A be an *arbitrary matrix* of $\mathbb{R}^{M\times N}$. Based on the description of the minimizers of G_{ℓ_0} given in [6], the two following theorems characterize the links between minimizers of G_{ℓ_0} and G_{CEL0}. Proofs can be found in [1].

Theorem 1: Let $d \in \mathbb{R}^M$ and $\lambda > 0$,

(i) the set of global minimizers of G_{ℓ_0} is included in the set of global minimizers of G_{CEL0},

$$\arg\min_{\mathrm{x}\in\mathbb{R}^N} \mathrm{G}_{\ell_0}(\mathrm{x}) \subseteq \arg\min_{\mathrm{x}\in\mathbb{R}^N} \mathrm{G}_{\text{CEL0}}(\mathrm{x}) \qquad (4)$$

(ii) conversely if $\hat{\mathrm{x}} \in \mathbb{R}^N$ is a global minimizer of G_{CEL0}, let $\hat{\mathrm{x}}^0$ be defined by

$$\forall\, i \in \mathbb{I}_N, \quad \hat{\mathrm{x}}_i^0 = \hat{\mathrm{x}}_i \mathbb{1}_{\{|\hat{\mathrm{x}}_i|\geq\frac{\sqrt{2\lambda}}{\|a_i\|}\}}\,, \qquad (5)$$

then $\hat{\mathrm{x}}^0$ is a global minimizer of G_{ℓ_0} and

$$\mathrm{G}_{\text{CEL0}}(\hat{\mathrm{x}}) = \mathrm{G}_{\text{CEL0}}(\hat{\mathrm{x}}^0) = \mathrm{G}_{\ell_0}(\hat{\mathrm{x}}^0)\,. \qquad (6)$$

Theorem 2: Let $d \in \mathbb{R}^M$, $\lambda > 0$, and G_{CEL0} have a local minimum (not global) at $\hat{\mathrm{x}} \in \mathbb{R}^N$. Then $\hat{\mathrm{x}}^0$ (defined by (5)) is a local minimizer (not global) of G_{ℓ_0} and (6) is verified.

Theorem 1 gives an "equivalence" between global minimizers of the two functionals while Theorem 2 partially extends this result to local minimizers: from all local minimizers of G_{CEL0} we can easily extract a local minimizer of G_{ℓ_0}. However the converse is false and we observed experimentally that an important amount of strict local minimizers of G_{ℓ_0} are not critical point of G_{CEL0} [1, §4.2]. In particular G_{CEL0} eliminates the strict local minimizers $\hat{\mathrm{x}}$ of G_{ℓ_0} such that $\sigma^-(\hat{\mathrm{x}}) \neq \emptyset$.

Note that, since from [6, Theorem 4.4 (i)], the set of global minimizers of G_{ℓ_0} is nonempty, Theorem 1 (i) ensures the existence of global minimizers for G_{CEL0}.

Then from Theorem 1 and 2 we conclude that it is preferable to address problem (1) by minimizing the *continuous* functional G_{CEL0} instead of G_{ℓ_0} since the global minimizers of G_{CEL0} *contain* those of G_{ℓ_0} and that G_{CEL0} has *"less" local minimizers* than G_{ℓ_0}.

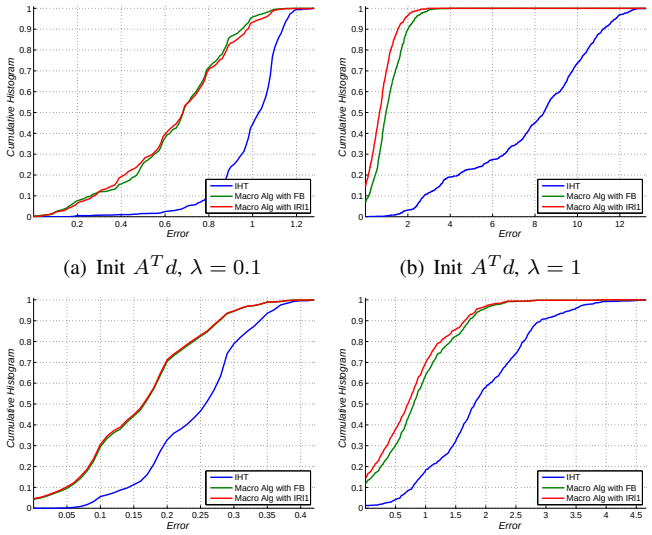
C. How to minimize G_{CEL0} ?

The continuity of G_{CEL0} allows to use nonsmooth nonconvex algorithms (*e.g.* [7], [8]) for minimizing G_{CEL0} and thus G_{ℓ_0}. Usually, such algorithms converge to a *critical point* of the minimized functional. Consequently, they cannot ensure the convergence to a minimizer of G_{CEL0}. However, the following lemma provides a relation between some critical points of G_{CEL0} and minimizers of G_{ℓ_0}.

Lemma 1: Let $\hat{\mathrm{x}} \in \mathbb{R}^N$ be a critical point of G_{CEL0} verifying $\sigma^-(\hat{\mathrm{x}}) = \emptyset$. Then it is a (local) minimizer of G_{ℓ_0} and $\mathrm{G}_{\text{CEL0}}(\hat{\mathrm{x}}) = \mathrm{G}_{\ell_0}(\hat{\mathrm{x}})$.

Therefore, from *any* state of the art algorithms verifying a *sufficient decrease condition* and the *convergence to a critical point of* G_{CEL0} (*e.g.* [7], [8]), we can define a *macro algorithm* [1, Algorithm 1] which adds an outer loop to move iteratively from a critical point of G_{CEL0} to another one while decreasing the cost function and converging to a point $\hat{\mathrm{x}}$ such that $\sigma^-(\hat{\mathrm{x}}) = \emptyset$ [1, Theorem 5.1]. From Lemma 1, such a point is a *(local) minimizer of* G_{ℓ_0}.

Using the *Iteratively Reweighted* ℓ_1 *(IRL1)* [7] or *Forward-Backward Splitting (FBS)* [8] as inner algorithm within the proposed macro algorithm, numerical experiments [1, §5.1] compare its performances with the *Iterative Hard Thresholding* (IHT) algorithm which is also ensured to converge to a (local) minimizer of problem (1) (see [5], [8]). A part of the experiments conducted in [1, §5.1] are reported on Fig 1. We can see the interesting behaviour of the proposed macro algorithm which converges to (local) minimizers of G_{ℓ_0} with a lower function value than those obtained with IHT. This shows that the macro algorithm is more "robust" against local minimizers of G_{ℓ_0} than IHT.



(a) Init $A^T d$, $\lambda = 0.1$ (b) Init $A^T d$, $\lambda = 1$

(c) Init ℓ_1, $\lambda = 0.1$ (d) Init ℓ_1, $\lambda = 1$

Fig. 1. Cumulative histograms of the error $|\mathrm{G}_{\ell_0}(\hat{\mathrm{x}}) - \mathrm{G}_{\ell_0}(\mathrm{x}^\star)|$ where $\hat{\mathrm{x}}$ is the estimated solution and $\mathrm{x}^\star$ is a global minimizer of G_{ℓ_0}. The histograms are computed from 1000 random matrices $A \in \mathbb{R}^{7\times 15}$ and $d \in R^7$ generated from an uniform distribution (the random seed is fixed at the same value for the four configurations (Init, λ) in order to generate the same sequence of problems). Two different initializations are considered: $\mathrm{x}^0 = A^T d$ (Init $A^T d$) and $\mathrm{x}^0 = \mathrm{x}_{\ell_1}$ (Init ℓ_1) the solution of the ℓ_1 relaxed problem. The experiment is repeated for two values of λ (0.1 and 1). For each configuration the estimation is performed using the IHT algorithm (blue) and the macro algorithm combined with IRL1 (red) or FBS (green).

The theoretical analysis of the proposed *tight continuous relaxation* G_{CEL0} of G_{ℓ_0} and the numerical experiments conducted on low dimensional examples are promising for the development of new algorithms to deal with problem (1) taking benefit from the nice properties of the G_{CEL0} functional.

REFERENCES

[1] E. Soubies, L. Blanc-Féraud, and G. Aubert, "A Continuous Exact ℓ_0 penalty (CEL0) for least squares regularized problems," *Preprint hal-01102492*, 2015.

[2] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[3] V. N. Temlyakov, "Greedy approximation," *Acta Numerica*, vol. 17, pp. 235–409, 2008.

[4] H. A. Le Thi, H. M. Le, and T. P. Dinh, "Feature selection in machine learning: an exact penalty approach using a difference of convex function algorithm," *Machine Learning*, pp. 1–24, 2014.

[5] T. Blumensath and M. E. Davies, "Iterative thresholding for sparse approximations," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 629–654, 2008.

[6] M. Nikolova, "Description of the minimizers of least squares regularized with ℓ_0-norm. Uniqueness of the global minimizer," *SIAM Journal on Imaging Sciences*, vol. 6, no. 2, pp. 904–937, 2013.

[7] P. Ochs, A. Dosovitskiy, T. Brox, and T. Pock, "On Iteratively Reweighted Algorithms for Nonsmooth Nonconvex Optimization in Computer Vision," *SIAM Journal on Imaging Sciences*, vol. 8, no. 1, pp. 331–372, 2015.

[8] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized gaussseidel methods," *Mathematical Programming*, vol. 137, no. 1-2, pp. 91–129, 2013.